Score-level Multi Cue Fusion for Sign Language Recognition

Çağrı Gökçe[1], Oğulcan Özdemir[1], Ahmet Alp Kındıroğlu[1], and Lale Akarun[1]

Bogaziçi University, Computer Engineering Department, Istanbul, Turkey
{cagri.gokce, ogulcan.ozdemir, alp.kindiroglu, akarun}@boun.edu.tr

Abstract. Sign Languages are expressed through hand and upper body gestures as well as facial expressions. Therefore, Sign Language Recognition (SLR) needs to focus on all such cues. Previous work uses handcrafted mechanisms or network aggregation to extract the different cue features to increase SLR performance, which is slow and involves complicated architectures. We propose a more straightforward approach which focuses on training separate cue models specializing on the dominant hand, both hands, face, and the upper body regions. We compare the performance of 3D Convolutional Neural Network (CNN) models specializing in these regions, combine them through score-level fusion, and use the weighted alternative. Our experimental results have shown the effectiveness of mixed convolutional models. Their fusion yields up to 19% accuracy improvement over the baseline using the full upper body. Furthermore, we include a discussion for fusion settings, which can be beneficial for future work on Sign Language Translation (SLT).

Keywords: Sign Language Recognition, Turkish Sign Language (TID), 3D Convolutional Neural Networks, Score-level Fusion

1 Introduction

Sign Language is the means of communication for the Deaf, and each Deaf culture has its sign language. Sign languages differ from the spoken language of the culture. Communication between the Deaf and the hearing impaired relies mostly on the Deaf individual learning the spoken language and using lipreading and written text to communicate: A huge and unfair burden on the Deaf. The reverse, teaching the general population at least some sign language may be more accessible, and there are available educational courses for such aim. However, gaining expertise in sign language is difficult, and the communication problem is still unsolved. Automatic interpretation of sign languages is a necessary step for not only enabling the human-computer interaction but also facilitating the communication between the Deaf and the hearing impaired individuals.

Automatic Sign Language Recognition (ASLR) refers to a broad field with different tasks, such as recognizing isolated sign glosses and continuous sign sentences. The objective of the ASLR system is to infer the meaning of the sign glosses or sentences and translate it to the spoken language. Recently, there has been increased progress in these efforts: Sign Language Translation (SLT) has become an active research problem for creating interactive sign language interfaces for the deaf [1, 2, 17]. A number of recent papers on the topic made use of neural network generated features. However, while the quality and representative power of these features in SLT are essential, and it is difficult to evaluate the representative potential of the elements in a pipeline setting where the overall system error is cumulative. For this reason, in this study, we aim to evaluate 3D Residual CNN Based Sign Language embeddings in terms of explanatory power in an Automatic Sign Language Recognition (ASLR) setting where temporal mix-up between signs and co-articulation is minimal. For the general case of Isolated SLR, the system aims to process a sign gloss and assign it to a single sign gloss label. In a limited context of supervised learning set-up, labels are glosses, which are transcription symbols assigned by sign language experts. There may be a single signer or multiple signers in communication; however, the ASLR system should be signer independent.

To convey the meaning of a performed sign gloss, Sign Languages use multiple channels, which are manifested as visual cues. We can classify these visual cues into two categories; (1) cues that are denoted as manual cues including hand shape and movement, and (2) cues that are non-manual features including facial expressions and upper body pose focusing on details without definitive large displacements.

Solving the problem of Isolated SLR requires specialized methods, divided into two categories. The first category is the handcrafted features, focusing on the video trajectories and flow maps [27, 18, 25]. The second set of methods includes machine learning algorithms and neural networks to improve classification performance [13, 23, 18]. 3D CNN models, a state-of-the-art deep neural network architecture, have proven successful in various video tasks [21, 22]. Li et al. [13] adopted the same architecture in the SLR and reported improved performance. However, Ozdemir et al. [18] provided the comparison of 3D CNN models and the handcrafted methods but have found 3D CNN's to be inferior to a state-of-the-art handcrafted human action recognition approach, Improved Dense Trajectories [25].

The aim of this work is to investigate why 3D CNN models may fail to show similar success in sign language recognition and to observe what modifications improve their performance. We hypothesize that the performance drop occurs because of the common practice of scaling images into smaller size and sampling frames [21, 22], due to computational requirements and difficulty of training bigger neural networks. One solution is handling the negative effect of the sampling by increasing the model complexity as in [5, 28, 11], yet this increases computational requirement. Instead, we firstly apply attentive data selection at the pre-processing phase by determining cues on SLR data. Secondly, we divide the problem into multiple cues and train different expert classifiers on each kind of dense feature. Thirdly, we refine the expert cue network knowledge into one result, by applying score-level fusion.

The paper organization is as follows. Sections 2 reviews related work, 3 explains the presented method, 4 presents the experimental results, 5 contains the analysis of experiments and 6 presents the conclusions.

2 Related Work

Sign Language Recognition (SLR) aims to infer meaning from a performed sign. In the sign classification task, an isolated sign is assigned a class label. A sign gloss, the written language counterpart of the performed sign, can be used as a mid-level or final stage label for such recognition in the supervised setting.

Sign Language Recognition is closely connected with video recognition or human action recognition methods, and similar architectures have been used for both. Two popular approaches to sign language representation uses handcrafted features and deep neural network based methods.

Prior to the performance leap achieved by neural networks, hand-crafted features were the best performing approach for representing human actions in a sequential video setting. For a two-frame dynamic flow map estimation, the Optical flow method is used to generate feature-level information. These features perform better representation than RGB image sequences in settings where motion is more indicative than shape [4]. While there existed numerous handcrafted feature extraction methods and their application to image sequences such as STIP [14] and spatio-temporal local binary patterns [26], state of the art performances with constructed features in action recognition and isolated sign language recognition were obtained using Improved Dense Trajectories [25, 18] which is an outlier independent trajectory-based motion specialized feature extractor.

Neural Network based methods focus on the convolutional architectures for the classification task. Simonyan et al. [19] use branched CNN architecture that splits the spatial and temporal data into each branch, achieving the latter using the optical flow map between frames. Tran et al. [21] use 3D convolutional kernels to build a 3D CNN variant to process video data in the end to end fashion.

One prerequisite for using deep neural networks is the presence of the large datasets with ground truth annotations. Recently, big-scale isolated sign language recognition datasets have become publicly available. Isolated SL datasets contain videos of a user performing a single gloss, usually a single word or a phrase. MS-ASL [23] is an American Isolated SL dataset, including 200 native performers performing more than a thousand word categories. WL-ASL [13] is a bigger dataset with two thousand word categories performed by one hundred people. For other languages, Chinese [27] and Turkish [18] are among the available datasets. Popular human activity recognition datasets [20, 12, 9, 10] are also used as an extra data and for finetuning in isolated SLR. Continuous SL datasets are acquired in a less controlled setting, where a user can perform longer sign sequences [7, 6].

SLR methods often use video pre-processing to reduce network bias and variance and to increase network performance. Random cropping is one of the popular spatial augmentation techniques when training CNNs. Since CNN variants have small input spatial resolution, e.g., 224 × 224 for the popular ResNet50 network [8], such methods increase the transitive invariance of the models by of processing different parts of the image in higher resolution compared to directly downsampling the whole image frame.

Temporal pre-processing techniques operate on the temporal dimension of the video data. The aim is to locate the dense temporal regions which have an increased likelihood of the action flow. In recent work, different approaches are applied for the temporal activity localization, e.g., exploiting both short term and long term samples [24], combining high and low-frequency learners [5], and detecting active window boundaries for the long sequences [16]. Our work differs by applying cue selection before training phase and combining the classifiers in the data augmentation level.

Combining the both pre-processing techniques allows an opportunity to exploit covariance between these spatio-temporal features. Spatio-temporal pre-processing can possibly improve the signal to noise ratio of the processed data when the region of interest is selected from dense regions. This process is shown to be beneficial on other video recognition tasks, e.g., when extracted through handcrafted methods such as optical flow [19], or directly through 3D CNNs [22]. In SLR, due to the nature of the task, sign language videos consists of the sparse hand and upper body movements as well as facial expressions. It is possible to use the domain-specific knowledge to exploit the spatio-temporal sampling using a guided pre-processing technique. Spatio-temporal multi cue networks [28] exploit spatial regions of interest by firstly using a branch to estimate the region of interest, then training different networks for each unit. However, applying sampling at the training phase becomes more computationally expensive and requires deeper architectures. Our score-level multi cue fusion approach address this problem and described in the next section.

3 Method

In this section, we describe our method. We firstly describe the mixed convolutional model, follow up with our Multi cue sampling process, and finally discuss the score-level fusion method.

3.1 3D Resnets with Mixed Convolutions

Mixed convolutional networks are 3D Convolutional networks [21], which use 3D convolutional kernels to process video data in an end to end fashion. Tran et al. [22] investigate the success of 3D CNNs and shares two effective variants with strong empirical results. The first is mixed convolutional networks, and the second is residual bottleneck based 2+1 convolutional networks.

Mixed CNN variant builds on the plain 2D residual networks, with the difference that the first layers are replaced with 3D convolutional kernels. First layers are capable of processing input video directly, and gradually lower the feature dimension into 2D, then feed into more efficient 2D convolutional last layers. A fully connected layer is employed after the final layer for the classification task.

Mixed Convolutional networks are denoted with MCx_n, where x is the number of 3D convolutional layer blocks, and n denotes the total number of layers. Following the baseline, we empirically experiment with different mixed convolutional variants and employ the $MC3_{18}$ variant of the mixed convolutional network.

3.2 Spatial and Temporal Sampling

The SLR task is conveyed through manual and non-manual cues. Information is conveyed through the shape and configuration of the hand, body, and face regions. The informative regions and intervals may be sampled with the help of a state-of-the-art pose estimation approach such as OpenPose [3]. Making use of pose estimation allows researchers to filter the entire frame by cropping specific regions according to keypoints, which are hand, face, and upper body keypoints in the case of SLR.

We would like to sample informative body regions to increase efficiency and to filter out noise. Our approach is two-fold. At the first stage, we design a SLR specific system by extracting the body, hand, and face regions using cropping in the spatial domain, as shown in Figure 1. To achieve this, we utilize the pose data as in [27, 18], to generate crops directly on the RGB image. In the second stage, we focus on the dense regions on the temporal domain. We define the active window as the temporal window where the active hand is moving. Then we filter out the sparse frames and only feed the network the frames in the active window, as shown in Figure 2.

Using Isolated data guarantees that the hand movement is in the middle of the temporal sequence. The following steps are used to extract the active window in the middle.

1. Use the moving hand detection framework in 4.2 to detect the active hand(s).
2. Define a selected hand as the active hand. If both hands are active, select the dominant hand.
3. For the selected hand, track hand movements using Euclidean distance. Keep the frame ids of the start of the first-hand movement and end of the last hand-movement.
4. Define two thresholds T_S, T_E. Filter the boundary regions from the start and end frame ids using corresponding thresholds defined earlier, and use extracted frames for the training.

In some videos, the movement is not in the middle of the video. We detect such exception cases by checking the position of the hand relative to the hip. We also filter out segments too short to be a sign.

3.3 Multi Cue Score Fusion

Extracting multiple cues from different settings allows each model to build expertise on each cue. Therefore, there is a need to distill the knowledge of each

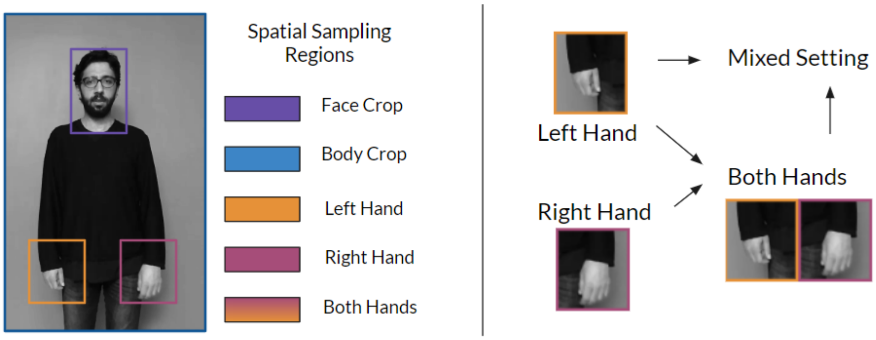

Fig. 1. Spatial Sampling operation is visualized. From left to right; cue regions selected for the process, and hand crop settings

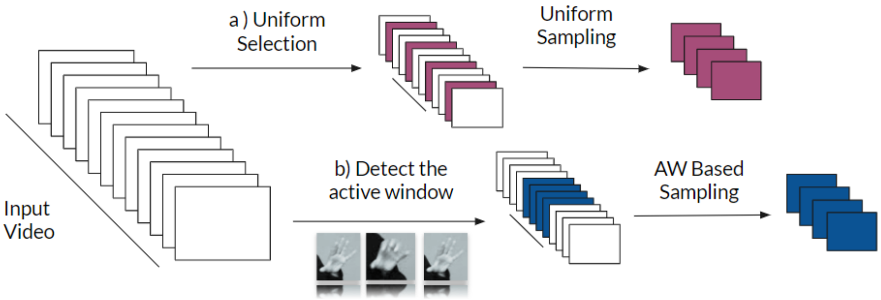

Fig. 2. Different temporal sampling operations are shown in the above figure. Selected frames are shown with color. Two branches represent uniform sampling and the Active Window Based Sampling Process

model by combining weak expert classifiers. Zhou et al. [28] experiments with distillation at the training time, by training a big scale model consisting of expert components. This has the drawback of increasing model complexity and training time. Simonyan et al. [19] combines different branches while training, but process the spatial and temporal branch separately at test time using a score fusion approach. They propose firstly direct score fusion via averaging through the network outputs and secondly, training a meta classifier above the extracted features. We follow the former score fusion approach since it has less model complexity and can achieve better run-time performance.

We experiment with two different multi cue fusion settings. First, we apply the averaging operation to the predicted scores of each cue network results. Secondly, we apply a weighted fusion, where each cue network is weighted by its validation set performance.

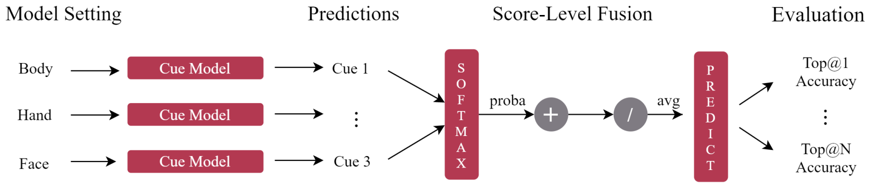

Fig. 3. Score-level multi cue fusion operation applied at the test time. Note that cue networks have different test weights even the architecture is same

4 Experiments

4.1 Experimental Setup

Dataset To achieve a competitive experiment setting, and to implement our proposal effectively, we have used a recently published Turkish isolated SLR dataset BosphorusSign22k [18]. The dataset contains six different native signers, performing 744 different word categories. Each category is labeled with a sign gloss, that describes the performed sign. The dataset contains about 22, 000 video clips. Authors also share 3D body pose keypoints in Kinectv2 format, and 2D body and hand keypoints obtained from OpenPose [3].

Evaluation Metric Following the work of Ozdemir et al. [18], we aim to compete on the sign language classification task. It is described as estimating the corresponding sign gloss for a given input video at test time, and scoring is evaluated in the accuracy of all of the test estimations. Out of all 6 performers, video clips of User 4 is defined as a test set, which is about 1/6 of the total dataset and it includes samples from all of the 744 classes.

Implementation Details Our experiment setting follows the baseline paper's[18] neural network based experiment setting. We apply provided preprocessing pipeline, resize the image into 640 × 360, crop the center square region then resize via bilinear interpolation to achieve 112 × 112 input resolution. We adopt the mixed convolutional MC3_18 CNN variant implemented in the torchvision library, start training process from the Kinetics [10] pretrained state, modify last 3 blocks, and apply uniform frame sampling. Our single model finetuning spans 48 hours with 32 batch size on Nvidia 1080 TI GPU.

Our replicated network resulted in 75.23% accuracy, which is over 3% lower than the reported 78.85% accuracy result. We suspect that the difference is caused by randomized states such as optimizer initialization and different hyperparameter choices such as the learning rate.

For pose estimation, have used OpenPose [3], an up-to-date 2D and 3D pose extraction framework. Openpose Body 25 estimation setting is used with 1920 × 1080 input video resolution, and extraction is applied with hand tracker active. Single frame processing took an average of 0.8 seconds on 1080 TI GPU. After the processing, the final keypoint file includes 18 keypoints for the body and 21 keypoints for each hand.

Table 1. Hand spatial sampling settings. First table represents hand activity distribution in the BS22k dataset. Second table represents test results of the different hand crop settings and resulting accuracy values

Distribution	Relative Frequency (%)
Both Active	**66.44**
Only Left Active	33.07
Only Right Active	0.40

Crop Setting	Accuracy(%)
Single Hand	79.13
Both Hands	85.81
Mixed	**86.25**

4.2 Experimental Results

Spatial sampling. Spatial sampling operation is applied through two phases. Firstly, the cue region is detected, cropped, and optionally concatenated in a multiple cue setting. Secondly, sampling is applied using bilinear interpolation.

Body Setting. Following the standard SLR pipeline, we apply cropping to the human body region before training. Refer to the Section 4.1 for the details.

Hand Setting. SLR work suggests the dominant hand, the most used hand, conveys the most information in communication. To detect the dominant hand in the BS22k dataset, we employ a hand motion tracking algorithm. This detection process is achieved by the following:

1. Detect the Thumb keypoints on each frame;
2. Define the first thumb keypoint on each hand as two anchors;
3. If the following thumb keypoint on the next frames has greater distance than threshold compared to the anchor, conclude the hand as moving.

The distance metric is selected as the Euclidian distance, and the threshold is selected as 150 pixels. Table 1 provides the detection results on moving hands-on BS22K dataset. We can conclude that BS22K dataset performers are using the left hand dominantly.

During signing, only one hand may be active, or both hands may be active. We have adopted three different policies: Firstly, the single cue setting is applied by selecting the dominant hand. The cropping procedure is applied to the 350 × 350 area around the keypoint #2 center. Secondly, both cue setting is applied by selecting both hands. Cropping is applied to 175 × 350 area around the Thumb keypoint, and each hand result is concatenated horizontally. Thirdly, the mixed setting uses a single cue approach when a single hand is active and uses both cue approach when both hands are active. All three policies are followed by the bilinear interpolation downsampling. Results are provided in the right-hand side of Table 1.

Face Setting. Sign language performers often have cues with face gestures and lip movements called mouthings that give a hint for the gloss of the sign. Since the size of the face is small, we crop the whole face. Openpose [3] provides the nose keypoint in the body keypoint set and this keypoint is used as a center to crop a 200 × 200 subsample.

Table 2. Classification accuracy results of the sampling and fusion settings. Three different settings are provided in the table. From left to right, (1) Single cue spatial sampling results, (2) Active Window Based Temporal Sampling applied to each crop, and (3) Spatial&Temporal settings are combined in one setting. Note that the bottom two rows include the fusion result of the above three models in each setting.

	Spatial		Temporal		S&T Combined	
Setting	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Body	75.73	93.88	81.83	96.02	86.91	98.17
Hand	86.25	97.61	88.70	97.59	91.73	98.72
Face	24.27	44.45	37.00	57.89	39.12	59.33
Fusion	90.63	98.92	93.88	**99.65**	94.47	**99.78**
Weighted Fusion	**92.18**	**99.27**	**94.03**	99.56	**94.94**	99.76

Score-Level Fusion We follow the insight that the different cue models can capture a different subset of features, which can lead to better results when combined effectively. Standard fusion is applied by averaging softmax outputs as in [19]. In the weighted setting, we have applied weights to each model proportional to their validation accuracy via standard multiplication. Table 2 provides the result of the fusion.

Temporal Sampling Standard SLR training pipeline involves using the standard uniform frame sampling. We propose the active window based temporal sampling, applied by firstly extracting the dense cue regions before applying the uniform selection. Active window is detected as the part that the active hand is moving and discard the rest of the temporal data.

We used double thresholding for finding the active window. We have found that the start threshold $T_S = 90$, and the end threshold $T_E = 50$ generates competitive empirical results. Using the temporal sampling framework, we have successfully segmented the active window for each video. Then, we applied uniform sampling along with our standard training pipeline. Results are provided in Table 2.

Spatio-Temporal Sampling We applied active window based temporal sampling on top of the spatial multi cue regions. The final spatio-temporal sampling framework has improved on both single settings. With the addition of score-level fusion, test accuracy reached to 94.94%, which is the best result in all proposed settings as seen in the Table 2.

Our best setting provides 16.09% improvement on our baseline neural network setting [18]. We also manage to beat their previous best hand-crafted state-of-the-art result with 6.41% accuracy rate. Whereas the previous best method uses more than ten times bigger input spatial resolution (640×640), complicated hand-crafted methods [25] and a second stage SVM classifier, our approach only contains a 3D CNN and a sampling pipeline. Comparison with the baseline results is shown in Table 3.

Table 3. Comparison with the baseline approaches IDT and MC3_18 model.

Method	Acc@1	Acc@5
Baseline IDT [18]	88.53	-
Baseline MC3_18 [18]	78.85	94.76
Weighted Fusion - S&T Combined	**94.94**	**99.76**

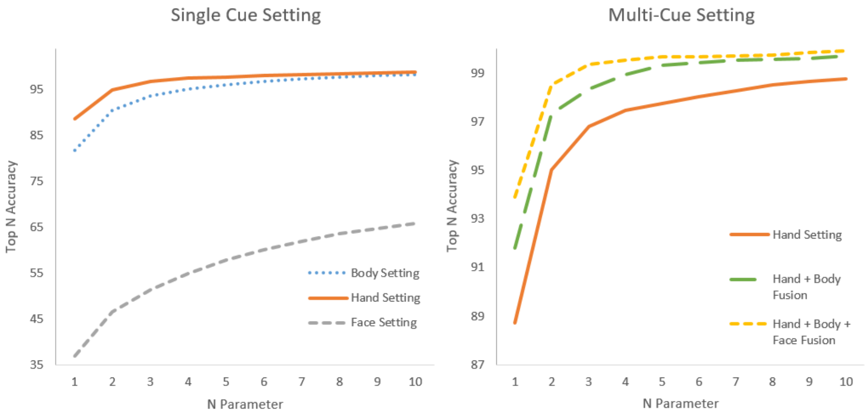

Fig. 4. Comparison of the models in the Top-N accuracy setting. The horizontal axis denotes the increasing N value, and the vertical axis denotes the accuracy value. First plot shows the Single Cue Setting comparison, and second plot shows the Multi Cue Setting additive comparison. Despite the difference in single setting performance, each cue boosts the fusion results.

5 Discussion and Analysis

Accuracy figures are not very informative when considering whether fusion will be beneficial. Top N Accuracy measures how often the top N ranks contain the correct class. In our experiments, we report Top-5 Accuracy along with Top-1 Accuracy. Top-N Accuracy results will increase with an increasing N, and are expected to be settled to 1 when N approaches to the maximum class number. Here, we provide graphs of Top-N accuracy versus N.

We share the Top-N accuracy graph in Figure 4. On the left-hand side, Top-N Accuraccies of the individual cues are reported. The hand cue performs the best, closely followed by the body cue. In both, there is a sharp increase between ranks 1 and 2. This shows that in a large number of cases, although the correct class fails to be predicted, it is the runner-up. This explains why the fusion is beneficial. Although the Top-N accuracy of the face cue is much lower, it is still beneficial for fusion.

Top-N accuracy of the muti-cue fusion is given in the right-hand side of Figure 4. We start by the hand model, then include the body model, and finally

Table 4. Effects of excluding individual cue units from the final fusion model. Using the different two cue settings and their performance, we infer to the excluded setting and its effect on the final mix.

Setting	Accuracy	Excluded Cue	Effect (%)
Body + Hand	91.80	Face	2.08
Body + Face	**84.66**	Hand	**9.22**
Hands + Face	88.70	Body	5.18

Table 5. F1-score comparison for the top ten sign glosses that hand sampling outperforms body sampling. (Sorted in the alphabetical order)

Sign Gloss	Hand	Body	Fusion	Sign Gloss	Hand	Body	Fusion
Aspirin	0.62	0.00	0.67	Internet_2	1.00	0.33	1.00
Deposit(v)_2	0.89	0.25	1.00	Noon	0.91	0.33	1.00
Exchange(v)	0.57	0.00	1.00	Shout(v)_2	0.91	0.33	0.91
Head	0.89	0.29	1.00	Sleep(v)	0.62	0.00	0.67
Identify(v)	0.89	0.00	1.00	Turn(v)	1.00	0.40	0.89

add the face model to the mix. This analysis allows us to see the cumulative progress over different fused models. We observe that the Top-2 Accuracy of hand alone is higher than Top 1 accuracy of both fusion settings. This shows why weighted fusion outperforms score fusion and shows that more advanced models can attain higher performance.

5.1 Spatial Ablation Study

Which cue benefits the fusion result the most? We have applied score fusion to all two-pair combinations of each cue setting. Using the ablation information, we can observe the effect of each cue to overall fusion. E.g., for finding the effect of the face model, we can subtract the Body+Hand setting result from the Body+Hand+Face setting. Table 4 shows the result of the using such calculation. It is seen that hands subunit has the most effect on fusion with 9.22% percent, followed by body model with 5.18%.

We have provided an analysis of the two most effective cues by comparing the gloss based performance. As a comparison metric, we adopted the F1-score, which should be more representative of false positives and false negatives, thus is more suitable for the gloss based evaluation.

Gloss Based Cue Comparison We share the top ten sign glosses that the hand cue model has a major advantage compared to the Body Cue model in Table 5.

We provide detailed analysis for the one example of the IDENTIFY(v) sign gloss. IDENTIFY(v) sign is performed by using only the left hand and touching

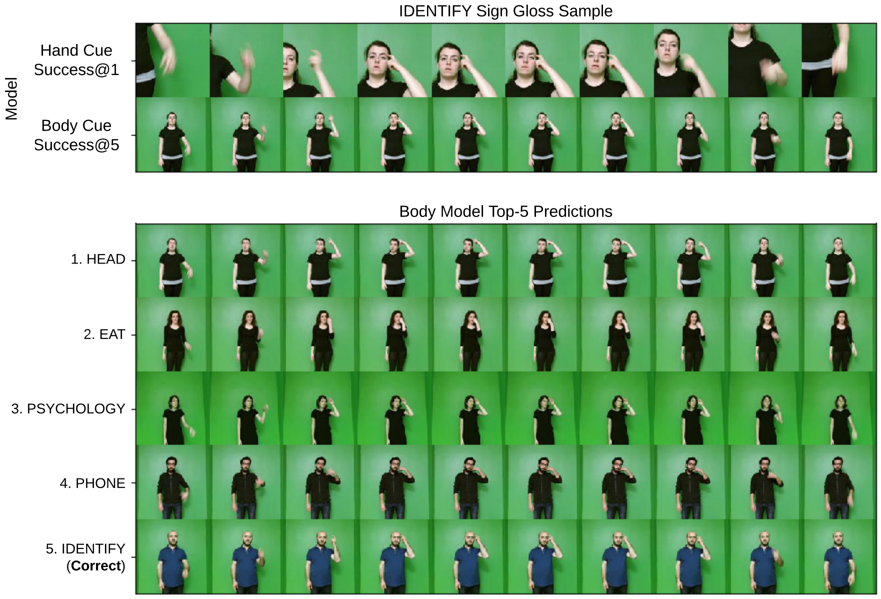

Fig. 5. Class confusions of IDENTIFY(v) sign gloss for the Body cue model

the head with the index finger, and the rest of the fingers are on the semi-open position. Whereas the hand cue model has the perfect score, the Body cue model only achieves success in the fifth guess.

Wrong predictions of the Body cue model are HEAD, EAT, PSYCHOLOGY, and PHONE sign glosses. We inspect each confusion as follows:

- **HEAD** sign differs from IDENTIFY(v) with the close position on all fingers other than the index finger.
- **EAT** sign is performed by moving the left hand close to the mouth and with all fingers are in a closed position.
- **PSYCHOLOGY** and **PHONE** sign glosses are performed with the left hand that and have open and semi-closed hand shapes, respectively.

By evaluating the confused cases, we conclude that the hand model has an advantage capturing hand shape information, possibly due to increased spatial resolution in the hand region.

Effect of the Score-Level Fusion. We share the Fusion Result of the Body and Hand cue models in Table 5. Data has shown that the fusion model successfully captures the hand cue features. The fusion model even outperforms both single cue models in 7 out of 10 glosses.

Table 6. Analysis of the temporal sampling based recognition approach with respect to signs with certain grammatical sign attributes: one-handed signs, two-handed signs, mono-morphemic signs, compound signs, and signs involving repetitive and circular movements, respectively

	Number of Classes with Selected Attribute								
	234	510	75	669	457	287	375	369	744
	One	Two	Circ.	Not	Rep.	Not	Mono	Comp.	All
	Handed	Handed		Circ.		Rep.			
Body	72.94	86.10	86.43	81.33	80.77	83.55	77.40	86.48	81.83
Hand	83.78	91.07	91.40	88.41	87.54	90.59	85.36	92.24	88.70
Face	45.33	33.01	29.41	37.82	36.39	37.99	33.79	40.52	37.00
Fusion	**91.82**	94.86	**96.15**	93.63	**93.45**	94.57	92.08	95.78	93.88
W.Fusion	90.87	**95.55**	95.70	**93.85**	93.37	**95.09**	**92.21**	**95.96**	**94.03**

5.2 Analysis of Method on Types of Gestures Recognized

To further analyze the types of signs where the presented method performs well and fails, we have labeled the 744 sign glosses in the dataset according to specific sign attributes. The sign classes are grouped into categories such as one-handed signs, two-handed signs, mono-morphemic signs, compound signs, and signs involving repetitive and circular movements of the hands.

Table 6 summarizes the analysis: The experiments are performed using temporal sampling with the best performing mixed convolution approach. Attribute-wise accuracy scores are calculated using the test set samples belonging to the classes containing the selected attributes. Overall, the accuracy scores in Table 6 demonstrate that for nearly all the subsets in the dataset, hand, body, and face-based features show consistency in their relative performance.

Looking at the results for different attributes one by one, we can see that signs involving two moving hands are better recognized than the one-handed sign glosses in the dataset. The performance difference can be explained by the fact that in one-handed signs, the weight of handshape may be more critical than the two-handed signs. The relative positioning and appearance of both hands, which is more apparent, may be easier to represent for the neural network.

Secondly, compound signs have a greater recognition accuracy than mono-morphemic signs (95.96% vs 92.21%). Considering the number of signing hands, the amount of additional information in the form of consecutive morphemes present in an isolated sign makes recognition easier, thus improving the performance system. From this result, we can infer that the method's representation power is higher when a sign is greater in length and contains different hand shape and position combinations.

Looking at repetitive gestures, we see a 1.6% improvement in accuracy when the signs do not contain repetitive hand gestures. The issue with repetitions, which we can attribute to this difference, is that the temporal and spatial forms of repetitions are more prone to differ between performances and users, in comparison to the static hand shape parts of the signs that follow specific rules.

Finally, we take a look at circular signs, which include circular hand and arm movements, which involve at least one entire rotation. These signs are dynamic signs where the hands do not stop while presenting a handshape. As these signs do not conform to the movement-hold phonological model of sign languages [15], representing them by choosing temporal frames is more complicated, reducing the effectiveness of keyframe based approaches [11]. Overall, the method performs well with circular signs, making fusion attempts with methods focusing more on the handshape of signs promising future leads.

6 Conclusion

In this paper, we proposed a score-level multi cue fusion approach for the Isolated SLR task. Unlike the previous work [18, 11], we focused on both spatial and temporal cues. We employed 3D Residual CNNs [22], and trained different models as an expert on the single cue. We distilled the expert knowledge using the weighted and unweighted score-Level fusion. In our experiments, we have seen that our approach has outperformed the baseline results on the Bosphorus-Sign22k Turkish Isolated SL dataset [18].

We have provided the single cue and multi cue Top-N accuracies to demonstrate incremental performance gain with each cue. Our gloss-level study shows that each cue model has specific expertise and provides an indispensable knowledge source to the fusion model. Our analysis of sign gloss attributes hints that the method performs better on temporally more complex signs with two-handed gestures, it shows comparatively worse on mono-morphemic gestures with a single hand. For that reason, the primary approach to improving performance lies in improving hand shape recognition. Possible strategies involve increasing model depth, finding better optimization techniques, or increasing the model input size. We hope that the following work will extend the SLR cues into other Sign Language problems, help progress in unresolved SL tasks such as translation, and help uncover language-independent cues.

References

1. Camgoz, N.C., Hadfield, S., Koller, O., Bowden, R., Ney, H.: Neural Sign Language Translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
2. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
3. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2019)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733 (July 2017)
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6201–6210 (2019)
6. Forster, J., Schmidt, C., Koller, O., Bellgardt, M., Ney, H.: Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC) (2014)
7. Hanke, T., König, L., Wagner, S., Matthes, S.: DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In: Proceedings of the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (2010)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, 2016. pp. 770–778 (June 2016)
9. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (June 2014)
10. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv:1705.06950 (2017)
11. Kındıroğlu, A.A., Özdemir, O., Akarun, L.: Temporal accumulative features for sign language recognition. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 1288–1297. IEEE (2019)
12. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: A large video database for human motion recognition. In: 2011 International Conference on Computer Vision. pp. 2556–2563. IEEE (November 2011)
13. Li, D., Opazo, C.R., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1448–1458 (2020)
14. Li, Y., Xia, R., Huang, Q., Xie, W., Li, X.: Survey of spatio-temporal interest point detection algorithms in video. IEEE Access **5**, 10323–10331 (2017)
15. Liddell, S.K., Johnson, R.E.: American sign language: The phonological base. Sign language studies **64**(1), 195–277 (1989)
16. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: Bmn: Boundary-matching network for temporal action proposal generation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3888–3897 (2019)
17. Orbay, A., Akarun, L.: Neural sign language translation by learning tokenization. arXiv preprint arXiv:2002.00479 (2020)

18. Özdemir, O., Kındıroğlu, A.A., Camgöz, N.C., Akarun, L.: Bosphorussign22k sign language recognition dataset. arXiv preprint arXiv:2004.01283 (2020)
19. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS. vol. 1, pp. 568–576. MIT Press (2014)
20. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402 (2012)
21. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV, 2015. pp. 4489–4497. IEEE (December 2015)
22. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
23. Vaezi Joze, H., Koller, O.: Ms-asl: A large-scale data set and benchmark for understanding american sign language. In: The British Machine Vision Conference (BMVC) (September 2019)
24. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1510–1517 (2017)
25. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: 2013 IEEE International Conference on Computer Vision. pp. 3551–3558. IEEE (2013)
26. Wang, Y., See, J., Phan, R.C.W., Oh, Y.H.: Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. PloS one **10**(5), e0124674 (2015)
27. Zhang, J., Zhou, W., Xie, C., Pu, J., Li, H.: Chinese sign language recognition with adaptive hmm. In: 2016 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2016)
28. Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for continuous sign language recognition. In: AAAI. pp. 13009–13016 (2020)